Teaching Business Statistics With Real Data To Undergraduates And The Use Of Technology In The Class Room

Rao Singamsetti, (E-mail: singamset@hartford.edu), University of Hartford

ABSTRACT

In this paper an attempt is made to highlight some issues of interpretation of statistical concepts and interpretation of results as taught in undergraduate Business statistics courses. The use of modern technology in the class room is shown to have increased the efficiency and the ease of learning and teaching in statistics. The importance of using real data in examples and exercises in class room teaching is emphasized.

INTRODUCTION

The main purpose of this paper is to address the fundamental issues of interpretation, data usage and efficient use of technology in teaching Business Statistics. These issues are extremely important in view of the fact that many of the undergraduate programs in the Business Schools do not require calculus course as a prerequisite to Business Statistics. The main reason to do away with calculus as a prerequisite is due to the fact that it fails to provide direct utility in understanding the concepts and applications in business. Moreover, the availability of the state- of- the- art computer technology facilitates broader coverage of topics as well as deemphasizing the calculus based techniques and computations. For example, computations on large scale data sets have become easier by the use of computer technology.

In section II , an attempt is made to present proper interpretation of several key concepts in Business Statistics. For example, the concept of confidence interval is essential in control charts and other applications. In section III, examples which can be worked out using computer technology and excel software efficiently in the class room are discussed. And, in section IV conclusions are presented.

SECTION II

Concepts

There are many ways to teach statistics. In Arts and Science Colleges, Statistics is taught as a branch of Applied Mathematics. Importance is given to definitions of concepts, derivation of formulas and proofs of lemmas and theorems. In Business Schools, emphasis is placed on the concepts, use of formulas without their derivations and practical applications in the areas of Accounting, Economics, Finance, Insurance, Marketing and Management. Use of computer technology in the class room made it easy to use the formulas. Computer software gives various kinds of probabilities, random samples from different populations, confidence intervals, information to be able to test hypotheses, correlations and regressions instantly. Therefore, what the Business student should learn in Business Statistics is to understand statistical concepts and use them, with the help of computer technology, in analyzing practical data and make appropriate inferences. In what follows, we will survey some important concepts where students find it difficult to grasp.

Probability

The concepts of probability, conditional, marginal and independence in the probability sense are often confusing to students. Probability concept is the law of large numbers. A practical example is a T.V. game show 'Let Us Make a Deal' in 1970's. In this game, a contestant is first given a choice of one of 3 doors. Behind one of the three doors there is a valuable item like a car and behind others there are goats or donkeys. When the contestant chooses one door, the host of the show reveals another door which has goat and asks the contestant if he/she changes the selection. Then the player may get the impression that by switching the chance of winning the car may improve from 1/3 originally to ½. But it is not correct. To start with there is 1/3 chance of choosing the desirable door. If the contestant chooses the wrong door initially which happens 2/3 of the time, switching becomes a winning strategy with the same probability 2/3. This paradox is well demonstrated by Java applet on the website (http://www.duke.edu./sites/java.html)

Sampling And Sampling Distributions

This is an important concept in inference in Statistics. Students are often confused between population and sample and the possibility of a number of possible samples of a given size from the same population. Construction of confidence interval for a parameter of the population and its interpretation is difficult for the students to understand. Students think that the confidence coefficient of say, 95%, means it is the probability that the population parameter lies within the lower and upper limits computed from the sample values. The correct interpretation should recognize the fact that the population parameter is an unknown constant and that the interval constructed from the sample is random. The lower and upper limits are computed from sample and hence the interval changes from sample to sample of the same size from the given population. Some intervals include the parameter and some don't. The confidence coefficient of 95% represents the proportion of intervals that include the parameter if we take unlimited number of samples from the population and construct intervals following a certain procedure. This concept is also well illustrated by a java applet at the website referred to above. We illustrate this concept in the next section using simulation with excel software package. Since simulation is done a limited number of times, the proportion of times the intervals cover the parameter may not equal the confidence coefficient but fluctuates around the confidence coefficient.

Hypotheses Testing

Another area where students find it difficult is to set up null and alternative hypothesis which is the first step in a five step process of testing of hypothesis on parameter/s of a population. Usually a claim made about the population parameter is considered as null hypothesis and it includes equality since this value is used in the computation of the value of the test statistic from the sample information. The word null, here implies that there is no difference between the stated value and the true value of the population parameter. This null hypothesis can be tested against a two-sided ($\neq$) or one sided ($>, <$) alternatives. Some times when the claim is not evident, the students are confused as to how to set up the null and alternative hypotheses. In such circumstances, as a rule, one should set up the alternative hypothesis first as ' what you want to show' (proposition to be established) based on the underlying theory and other pertinent information and opposite to this is the null hypothesis which should include equality. When the alternative is $<$, the rejection region is on the left extreme, when the alternative is $>$, the rejection region is on the right extreme and when the alternative is $\neq$ the rejection region is on both extremes. Rejection region is the portion of the sampling that is not consistent with the null hypothesis.

Students are also confused about setting up the decision rule using the value of the computed test statistic and the critical value found from probability tables. Equivalent decision rule can also be formulated using the p-value, that is, the probability of getting as extreme a statistic value as the one computed from sample or more extreme. If this p-value is less than the chosen value of α then reject the null hypothesis, otherwise do not reject the null hypothesis. These two ways of arriving at the same decision in testing of hypothesis can be done using Microsoft excel software package. The t test for means using two samples is presented in the next section.

Testing for independence of row and column classifications in a contingency table using 'Chitest' in the 'functions' category of excel presents some difficulty because one has to compute the expected frequencies. Excel gives p-value of Chi-square test statistic only. If you desire value of the Chi-square statistic you have to use 'Chiinv' function separately. This is not only time taking but cumbersome too. Other soft-ware packages like Minitab do not require computation of expected frequencies, but gives additional output, given the observed frequencies.

Correlation

Usually this is the pre-final topic that is covered in Business Statistics course for the undergraduates. Correlation is defined between two variables (X and Y). A sample of n observations can be plotted as a scatter diagram using 'Chart Wizard' menu button in excel. Pearson Correlation Coefficient has two aspects. One is sign indicating the relationship, + for positive or direct relationship and – for negative or inverse relationship between the two variables. The second aspect is the absolute numeric value indicating the strength of linear relationship between the two variables on a scale of 0 to 1. The value of zero represents absence of linearity and the value of one represents perfect linearity. Statistical independence of the two variables implies zero correlation but the converse is not true. If all the points in the scatter diagram lie on a straight line parallel to either X or Y axis, then correlation coefficient is undefined. The formula assumes 0/0 indicating there are no two variables. The correlation coefficient is symmetric between X and Y in the sense that it does not recognize the cause and effect relationship between the variables. This kind of discussion of correlation coefficient is absent in many text books on Business Statistics.

Regression

A discussion of bi-variate random variable can start with a contingency table and chi square statistic to test for independence of row and column classifications. Rejection of this null hypothesis of independence leads to the conclusion that row and column variables are associated and the table will be meaningful. But chisquare test gives neither the nature of the association nor strength of association. An improvement over this is correlation coefficient. Though correlation coefficient gives the direction and strength of linear relationship it does not indicate causal relationship between the two variables because it is symmetric. Further improvement over this is regression.

Regression is the final topic that is usually covered in Business statistics. It gives how the average value of the dependent variable is linearly related to independent or causal variable/s. This hypothesized relationship in the population, under certain assumptions, is estimated from a sample of observations. Estimation and analysis can be done easily using regression program (used for both simple and multiple regression) of 'Data Analysis' in excel. The designation of causal (X) and effect (Y) variables is to be specified by theories in economics or other disciplines.

An example and some points of clarification are presented in the next section.

SECTION III SOME EXAMPLES ILLUSTRATING THE CONCEPTS AND THEIR COMPUTATION USING MICROSOFT EXCEL SOFTWARE PACKAGE.

Confidence Interval

Using 'Random Number Generation' program in 'Data Analysis' package of Excel, ten random samples are drawn from a normal population with mean 50 and standard deviation 10. Lower limit (L limit) and Upper limit (U limit) for 95% confidence are computed using the formula, namely, sample mean +/_ t value corresponding to 9 d.f. times standard error of the mean assuming population standard deviation is not known. The results are presented below in Table 1.

Table 1. 95% Confidence Intervals For Ten Samples

Sample									L limit	U limit	
1	54.280	38.152	47.249	52.939	49.966	52.999	39.874	56.866	43.332	54.154	yes
2	38.785	51.645	59.157	38.098	45.906	64.692	35.188	46.976	38.801	57.707	yes
3	47.814	47.883	52.507	50.214	58.917	35.515	48.395	47.787	43.179	54.624	yes
4	45.261	50.638	46.287	45.751	22.578	53.157	32.996	69.079	34.223	58.200	yes
5	25.146	58.430	40.976	61.008	48.919	62.440	36.063	37.732	35.011	58.848	yes
6	49.693	56.302	32.586	62.813	56.279	44.321	36.702	64.974	40.574	61.241	yes
7	42.578	51.862	43.763	42.455	36.232	43.713	34.249	46.296	38.039	47.627	no
8	62.533	36.602	42.476	54.387	42.957	67.236	52.727	49.532	42.356	57.829	yes
9	41.971	22.306	36.002	58.370	46.245	46.711	46.075	52.676	34.652	53.907	yes
10	66.608	26.046	54.771	42.676	32.278	43.036	45.267	60.118	34.925	53.336	yes
Only one out of 10 intervals constructed does not contain the true mean of 50. That is 90% of intervals cover the true population mean even though they are 95% confidence intervals. The discrepancy is due to limited number of samples											

Testing Of Equality Of Means

The following examples based on real data are provided by my colleague, Dr. B.Kolluri.

DAILY CALORIES INTAKE PER CAPITA FOR CANADA AND USA

OECD (Organization for Economic Co-operation and Development) has been publishing health statistics since mid 1980s. According to the data published in 2003 covering the period 1981-2000, the mean and the standard deviation of daily per capita calories intake for Canada were 3041.65 calories and 87.64 calories, respectively. For the same period, the US data produced a mean value of 3488.85 and standard deviation of 184.42 calories

Data

Total Calories Intake For Canada And USA - Calories /Capita/Day

	Canada	United States
1981	2875	3209
1982	2926	3179
1983	2885	3217
1984	2971	3260
1985	3025	3364
1986	3063	3336
1987	3112	3437
1988	2986	3444
1989	3010	3425
1990	2995	3486
1991	3007	3513
1992	3057	3549
1993	3025	3592
1994	3133	3654
1995	3088	3597
1996	3079	3616
1997	3098	3682
1998	3157	3698
1999	3167	3747
2000	3174	3772

Source: 2003 OECD Health Data

From the data given above we can test the equality of mean intake of calories per capita per day between U.S.A and Canada against one sided (can we conclude that the per capita calorie intake per day for U.S.A. is more than that of Canada) alternative suggested by the sample information or two sided alternative. Since we have small samples we can use t-test (assuming normal populations and independence of samples) by using 't-test: Two Sample Assuming Unequal Variances' program in 'Data Analysis' under 'Tools' menu of Excel. The results are given in Table 2 below. The easy way to test one sided alternative is to compare the p-value for one tail which is 0 with α = say 0.05 and decide to reject the null hypothesis of equality of means in favor of the alternative. Similarly, we can reject the null hypothesis of equality of means against the two sided alternative by comparing the p-value for two-tail and α value.

Table 2 Results Of Testing Of Equality Of Means

	Canada	**United States**
Mean	3041.65	3488.85
Variance	7681.397	34009.4
Observations	20	20
Hypothesized Mean Difference	0	
df	27	
t Stat	-9.79483	
P(T<=t) one-tail	1.11E-10	
t Critical one-tail	1.703288	
P(T<=t) two-tail	2.21E-10	
t Critical two-tail	2.051829	

t-Test: Two-Sample Assuming Unequal Variances

In the example below, testing for independence of ethnicity and educational attainment using chitest function in excel software is illustrated.

Data

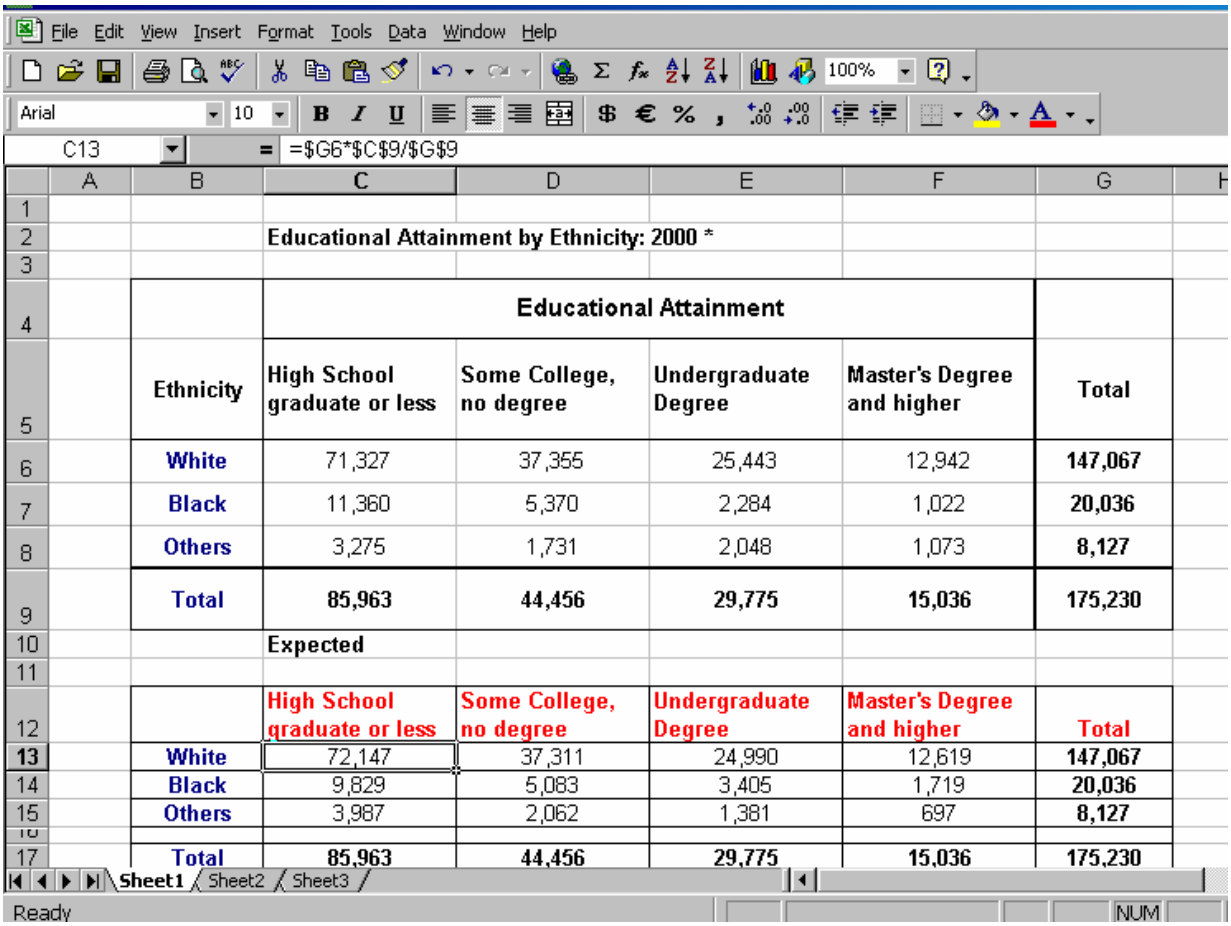
U.S. Census Bureau reported in the 2002 Statistical Abstract of the United States "**educational attainment by ethnicity"** figures (in thousands) for the year of 2000 as follows:

Educational Attainment by Ethnicity: 2000*

Ethnicity	**Educational Attainment**				
	High School graduate or less	**Some College, no degree**	**Undergraduate Degree**	**Master's Degree and higher**	**Total**
White	71,327	37,355	25,443	12,942	**147,067**
Black	11,360	5,370	2,284	1,022	**20,036**
Others	3,275	1,731	2,048	1,073	**8,127**
Total	**85,963**	**44,456**	**29,775**	**15,036**	**175,230**

Source: U. S. Census Bureau, Statistical Abstract of the United States: 2002 Table No. 210
* For persons 25 years old and over.

According to the figures in the data table above, can we conclude that the educational attainment and ethnic background are associated? Use a 0.01 significance level.



File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

C13 = =$G6*$C$9/$G$9

Educational Attainment by Ethnicity: 2000 *

	Educational Attainment				
Ethnicity	High School graduate or less	Some College, no degree	Undergraduate Degree	Master's Degree and higher	Total
White	71,327	37,355	25,443	12,942	**147,067**
Black	11,360	5,370	2,284	1,022	**20,036**
Others	3,275	1,731	2,048	1,073	**8,127**
Total	**85,963**	**44,456**	**29,775**	**15,036**	**175,230**

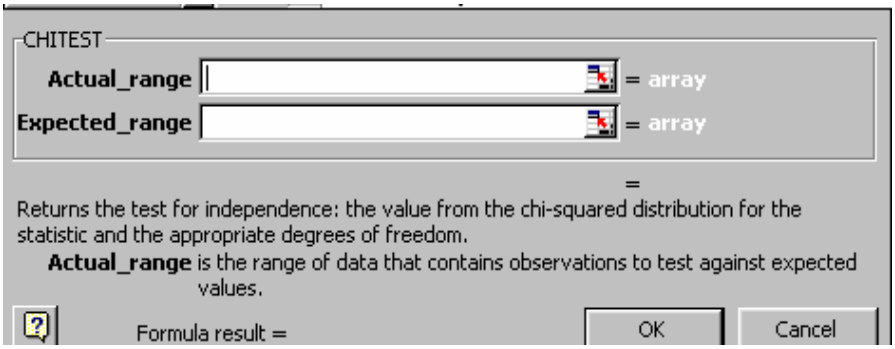
Expected

	High School graduate or less	Some College, no degree	Undergraduate Degree	Master's Degree and higher	Total
White	72,147	37,311	24,990	12,619	**147,067**
Black	9,829	5,083	3,405	1,719	**20,036**
Others	3,987	2,062	1,381	697	**8,127**
Total	**85,963**	**44,456**	**29,775**	**15,036**	**175,230**

Sheet1 Sheet2 Sheet3

Ready NUM

Given the data , we have to open a spread sheet in excel and enter the observed frequencies (actual range) and **compute** the expected frequencies(expected range) in a similar table as shown above. Then we can use chitest function in excel and complete the dialog box below to get the p-value.



CHITEST

Actual_range = array

Expected_range = array

=

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Actual_range is the range of data that contains observations to test against expected values.

Formula result = OK Cancel

If the generated p- value is less than the α value, we reject the null hypothesis of independence of ethnicity and educational attainment and conclude there is evidence to show that they are associated. This example shows the short coming of excel software in that it does not simplify the calculations as some other software packages do. It raises the question of suitability of this software in the class room for Business Statistics.

In the final example below we demonstrate the use of regression program in excel and indicate some problem areas in inferring from the excel output.

Data

Year	X (inflation)	Y (Total return %)
1926	-1.49	11.62
1927	-2.08	37.49
1928	-0.97	43.61
1929	0.2	-8.42
1930	-6.03	-24.9
1931	-9.52	-43.34
1932	-10.3	-8.19
1933	0.51	53.99
1934	2.03	-1.44
1935	2.99	47.67
1936	1.21	33.92
1937	3.1	-35.03
1938	-2.78	31.12
1939	-0.48	-0.41
1940	0.96	-9.78
1941	9.72	-11.59
1942	9.29	20.34
1943	3.16	25.9
1944	2.11	19.75
1945	2.25	36.44
1946	18.16	-8.07
1947	9.01	5.71
1948	2.71	5.5
1949	-1.8	18.79
1950	5.79	31.71
1951	5.87	24.02
1952	0.88	18.37
1953	0.62	-0.99
1954	-0.5	52.62
1955	0.37	31.56
1956	2.86	6.56
1957	3.02	-10.78
1958	1.76	43.36
1959	1.5	11.96
1960	1.48	0.47
1961	0.67	26.89
1962	1.22	-8.73
1963	1.65	22.8
1964	1.19	16.48
1965	1.92	12.45
1966	3.35	-10.06
1967	3.04	23.98

Year	X (inflation)	Y (Total return %)
1968	4.72	11.06
1969	6.11	-8.5
1970	5.49	4.01
1971	3.36	14.31
1972	3.41	18.98
1973	8.8	-14.66
1974	12.2	-26.47
1975	7.01	37.2
1976	4.81	23.84
1977	6.77	-7.18
1978	9.03	6.56
1979	13.31	18.44
1980	12.4	32.42
1981	8.94	-4.91
1982	3.87	21.41
1983	3.8	22.51
1984	3.95	6.27
1985	3.77	32.16
1986	1.13	18.47
1987	4.41	5.23
1988	4.42	16.81
1989	4.65	31.49
1990	6.11	-3.17
1991	3.06	30.55
1992	2.9	7.67
1993	2.75	9.99
1994	2.67	1.31
1995	2.54	37.43
1996	3.32	23.07
1997	1.7	33.36
1998	1.61	28.58
1999	2.68	21.04
2000	3.39	-9.11
2001	1.55	-11.88

Source: Ibbotson Associates – Valuation Edition Yearbook 2002

The excel regression output is given below. In the regression statistics part of the output if you take the square root of R Square you will get multiple R. But this is not the usual Pearson Correlation coefficient. It is the correlation between the dependent variable and the predicted average value of y, which is always positive (even in the case of multiple regression). In this simple regression case if you want to get the Pearson Correlation coefficient between X and Y, you have to attach a sign for the multiple R. Here it will have the same sign as the slope coefficient for X, that is, negative sign. So the correlation coefficient here is – 0.022954.

The main issue in using software packages such as excel is its limited scope. For example, the issue of stationarity in the time series is not addressed. Although it appears to be basic issue, the standard estimation and statistical test procedures are highly inappropriate, and even invalid, when the variables involved are non-stationary. Non stationary variables do not have constant means and variances and therefore violate the standard assumptions of regression model. For example, the excel printout shows insignificant relation between common stock returns and inflation. This is contrary to the common knowledge in the literature of a significant negative relation between these two variables. This might be the result of using a non-stationary data with the excel software. Excel also does not adjust for any possible auto correlation in the disturbance term. These limitations clearly indicate the need for an appropriate software to teach undergraduate Business Statistics.

Summary Output

Regression Statistics	
Multiple R	0.022954
R Square	0.000527
Adjusted R Square	-0.01298
Standard Error	20.369
Observations	76

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16.18500055	16.185	0.0390	0.843970831
Residual	74	30702.30964	414.8961		
Total	75	30718.49464			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.98566	2.879797711	4.509226	2.39E-05	7.247535625	18.72378137
X (inflation)	-0.1056	0.534681042	-0.19751	0.843971	-1.170979628	0.959771087

SECTION IV: CONCLUSIONS

An attempt is made in this paper to address the fundamental issues of interpretation of statistical concepts, usage of appropriate data and efficient use of technology in teaching undergraduate business statistics. The important concepts such as sampling distributions, confidence intervals, formulation of null and alternative hypothesis, correlation and regression are presented in a rigorous manner and their proper interpretations are discussed. Often most of the textbooks use made up data without any relevance to the reality. Several real data examples are used to present results using excel, and to draw appropriate inferences. The use of excel and its limitations are presented through four major project type examples. Briefly the following are the findings.

1. It is extremely important to emphasize interpretations with appropriate examples as demonstrated in the paper. This is especially important in a course like business statistics where there is less emphasis on theory and more emphasis on business applications.
2. A special effort is made in this paper to use the real data in examples and project type computer exercises to discuss the software results and draw inferences. This is essential to stimulate student interest in business statistics at undergraduate level.
3. The limitations of using excel as a statistical software are presented. This suggests that it is possible while the excel serves the basic need, other software such as MINITAB as a next higher level might be more appropriate. It is to be noted that SAS, TSP and others might be too advanced at the undergraduate level in Business schools.

Hopefully the issues presented in this paper stimulate discussion and analysis and lead to improved pedagogy and learning by students.

REFERENCES

1. Becker, E.W., Greene, W.H. (2001) Teaching Statistics and Econometrics to Undergraduates. *The Journal of Economic Perspectives*, Fall (15), pp.169-182.

NOTES